A stochastic approach to multi-gene expression dynamics

T. Ochiai*, J.C. Nacher*, T. Akutsu

February 9, 2008

Bioinformatics Center, Institute for Chemical Research, Kyoto University,

Uji, 611-0011, Japan

PACS number : 89.75.-k, 87.14.Gg, 87.15.Aa, 87.15.Vv

Keywords : Gene expression, Markov property, stochastic process.

Abstract

In the last years, tens of thousands gene expression profiles for cells of several organisms have been monitored. Gene expression is a complex transcriptional process where *mRNA* molecules are translated into proteins, which control most of the cell functions. In this process, the correlation among genes is crucial to determine the specific functions of genes. Here, we propose a novel multi-dimensional stochastic approach to deal with the gene correlation phenomena. Interestingly, our stochastic framework suggests that the study of the gene correlation requires only one theoretical assumption -*Markov property*- and the experimental transition probability, which characterizes the gene correlation system. Finally, a gene expression experiment is proposed for future applications of the model.

*These authors contributed equally to this work.
Corresponding authors: ochiai@kuicr.kyoto-u.ac.jp, nacher@kuicr.kyoto-u.ac.jp

1 Introduction

In a living organism or cell the collective behavior of thousands of genes and their products (for example mRNA and proteins) are embedded in a complex architecture that creates the mystery of life. More than 10 years ago, methods in molecular biology worked on a "*one gene- one experiment*" framework, meaning that the throughput is constrained to only one gene and therefore, the whole image of gene function is difficult to visualize. An emerging technology called *DNA* microarray/GeneChips [1, 2] appeared in the recent years, attracting the interests among biologists, computer scientists, mathematicians and physicists. This technology allows us to monitor the whole transcribed genome on a single chip and offers the possibility to capture the correlations among thousands of expressed genes simultaneously.

In order to understand how the organism works, it is necessary to know *which* genes are expressed, *when* they are expressed and *how fast* they do. Gene expression is regulated by means of the gene regulation architecture system of the cells, which involves network of interactions among *DNA*, *mRNA*, proteins and hundreds of small ubiquitous molecules. These interactions involve many elements and different and complex mechanisms, therefore an intuitive understanding of the underlying dynamics is not easy to obtain. This is also true for the issue of gene correlation, which is the main aim of this letter. In particular, although many approaches and techniques, as for example Boolean networks, graph theory and control theory, have been used successfully in many cases, they are still far to achieve a general description of the dynamics of the regulation and correlation among genes.

To shed light on this issue, here we propose a new theoretical model to deal with multi-gene correlation dynamics based on only one assumption: *Markov property* . Our approach will use the most general multi-variate stochastic process in order to obtain predictions about the correlations among genes.

In a previous work [3], we proposed a constructive approach to gene expression dynamics, which re-builds the scale-free organization of genes (i.e., expression level k decays as a power-law $k^{-\gamma}$ [4, 5]) observed in recent experiments [6, 7, 8]. There, we proposed a stochastic approach by assuming the Markov property and by using the observed experimental transition probability data, which characterize the gene expression system. Although our companion paper [3] succeeded to re-build the scale-free distribution, it may not provide much information about gene correlation phenomena because by construction it is one gene approach-like. Therefore, here our aim is to

exploit the novelty of our previous constructive approach by extending that one-dimensional model, to a multi-dimensional analysis (i.e., multi-gene correlation).

Our approach is based on two fundamental aspects: Markov property and stochastic process.

Markov property. If we say that the system has a Markov property, we mean that the future is governed by the present and does not depend on the past. Our model assumes Markov property to describe the multi-gene expression dynamics. Certainly, the living organisms are systems with long term memory, and they are complex systems with large number of elements and interactions. However, from a physical point of view, although we may not know all the variables in real situations, it may be enough to find a reduced number of variables whose behavior in time can be described as a Markovian process. Therefore, in our study we may assume that the most relevant degree of freedom of the system is the gene expression level, and consequently, our gene system has the Markov property.

Stochastic process. Although many complex systems may be governed by non-stochastic processes, in the gene expression problem the random variation is reasonable, plays a relevant role in cellular process, and furthermore stochastic noise have recently been measured and studied theoretically [9, 10, 11, 12]. For example, the expression level of thousands genes is very low, which creates intrinsic uncertainties in the number of expressed genes in the cells [6]. Furthermore, we can even distinguish between inherent stochasticity (*intrinsic noise*) and external stochasticity (*extrinsic noise*). While the origin of the first one are the biochemical processes, and motivates that two identical genes become uncorrelated due to that randomness, the second one represents sources of extrinsic noise, which change from cell to cell (i.e., fluctuations in elements among cells) [9, 13]. Moreover, the number of molecules which are involved in signal transduction pathways fluctuates from 10^2 to 10^4. Therefore, the randomness connected with elementary molecular interactions and their amplification in the signaling cascade generates significant spatio-temporal noise. Therefore, the stochastic approach is justified, and it seems more appropriate and plausible than a deterministic approach. Finally, it is also worth reminding that the current experimental techniques also provide an additional source of fluctuation, which come from the ubiquitous instrumental noise (which may be around 30% or more) from chip to chip with the current GeneChips technologies.

On the other hand, one drawback of our approach is that the current ex-

isting experimental data of gene expression time series gene lacks of enough statistics. For example, in yeast [14] and human [15] organism experiments, they analyzed the fluctuations in time of many genes simultaneously (more than 30.000 in the case of human organism) by carrying out only *one* experiment

However, in order to have enough statistics, we believe that by using many experiments of gene chips (i.e., *many* experiments measure repeatedly fluctuations of *many genes* in time under the same conditions), we may achieve a better understanding of the global nature and dynamics of gene correlation. Therefore, the theoretical approach proposed in this letter may be a useful guideline for such kind of experiments, and moreover may encourage them.

The paper is organized as follows. Section 2 describes the theoretical background, explains our proposed model for multi-gene correlations and presents the results of our simulated data. Section 3 explains the experimental proposal compatible with our theoretical study, and finally Section 4 presents the conclusions.

2 Methods and Result

2.1 Methods

2.1.1 Markov property and Differential Chapman-Kolmogorov Equation

Markov property. We use multi-dimensional stochastic process for describing the multi-gene correlation dynamics [16, 17, 18, 19]. Let $\{\mathbf{X}_t = X_t^1, \cdots, X_t^N), 0 \leq t < \infty\}$ be a multi-dimensional stochastic process. For $(t_n > \cdots > t_0)$, the conditional probability density function

$$p(\mathbf{x}_n, t_n|\mathbf{x}_{n-1}, t_{n-1}; \cdots; \mathbf{x}_0, t_0) = p(\mathbf{X}_{t_n} = \mathbf{x}_n|\mathbf{X}_{t_{n-1}} = \mathbf{x}_{n-1}; \cdots; \mathbf{X}_{t_0} = \mathbf{x}_0)$$

is defined as usual manner, where $\mathbf{x} = (x^1, \cdots, x^N)$ denotes the N dimensional vector. It is said that a multi-dimensional stochastic process has "Markov property", when the condition

$$p(\mathbf{x}_n, t_n|\mathbf{x}_{n-1}, t_{n-1}; \cdots; \mathbf{x}_0, t_0) = p(\mathbf{x}_n, t_n|\mathbf{x}_{n-1}, t_{n-1}) \tag{1}$$

holds for arbitrary $t_n > \cdots > t_0$. In what follows, we assume that the probability density $p(\mathbf{x}, t|\mathbf{x}_0, t_0)$ has the time translation invariance $p(\mathbf{x}, t|\mathbf{x}_0, t_0) = p(\mathbf{x}, t + a|\mathbf{x}_0, t_0 + a)$ for arbitrary a.

Our only one assumption is that *the multi-dimensional correlation dynamics of gene expression obeys the Markov property*. More precisely, we assume that the expression levels of each gene are denoted by the multi-dimensional stochastic process with Markov property $\mathbf{X}_t$.

Differential Chapman-Kolmogorov equation. For the matter of convenience, we write $p(\mathbf{x}, t)$ for $p(\mathbf{x}, t|\mathbf{x}_0, t_0)$. Then, if the multi-dimensional stochastic process has the Markov property (Eq. (1)), the conditional probability density function $p(\mathbf{x}, t|\mathbf{x}_0, t_0)$ obeys the Differential Chapman-Kolmogorov equation valid for a system composed of N genes and reads as follows:

$$\begin{aligned}\frac{\partial p(\mathbf{x},t)}{\partial t} &= -\sum_{i=1}^{N}\frac{\partial}{\partial x^i}\{a^i(\mathbf{x})p(\mathbf{x},t)\} + \frac{1}{2}\sum_{i,j=1}^{N}\frac{\partial^2}{\partial x^i x^j}\{b^{ij}(\mathbf{x})p(\mathbf{x},t)\} \\ &\quad + \int d\mathbf{y}[W(\mathbf{x}|\mathbf{y},t)p(\mathbf{y},t) - W(\mathbf{y}|\mathbf{x},t)p(\mathbf{x},t)], \qquad (2)\end{aligned}$$

where the *drift* term $a^i(\mathbf{x})$ is given by

$$\lim_{\epsilon\to 0}\frac{1}{\epsilon}\int_{|\mathbf{y}-\mathbf{x}|<\delta}(y^i - x^i)T_\epsilon(\mathbf{y},\mathbf{x})d\mathbf{y} = a^i(\mathbf{x}) + O(\delta), \qquad (3)$$

and the *diffusion* matrix $b^{ij}(\mathbf{x})$ reads as

$$\lim_{\epsilon\to 0}\frac{1}{\epsilon}\int_{|\mathbf{y}-\mathbf{x}|<\delta}(y^i - x^i)(y^j - x^j)T_\epsilon(\mathbf{y},\mathbf{x})d\mathbf{y} = b^{ij}(\mathbf{x}) + O(\delta), \qquad (4)$$

and *jump* term $W(\mathbf{y}|\mathbf{x}, t)$ is given by [1]

$$W(\mathbf{y}|\mathbf{x},t) = \lim_{\epsilon\to 0}T_\epsilon(\mathbf{y},\mathbf{x})/\epsilon. \qquad (5)$$

Here $T_\epsilon(\mathbf{y}, \mathbf{x})$ is an Instantaneous Transition Probability (ITP) defined by $T_\epsilon(\mathbf{y}, \mathbf{x}) = p(\mathbf{y}, t + \epsilon|\mathbf{x}, t)$ for sufficiently small ϵ.

In the context of gene expression level, Eq. (2) represents the dynamics of N *mRNA* molecules (i.e., gene expression) in the cell. By using this equation, we can study the dynamics and correlation between genes i and j.

Next, we will explain three important processes of the Differential Chapman-Kolmogorov equation (2) in the following paragraph.

[1]Here we remark that although $W(\mathbf{x}|\mathbf{y}, t) = 0$ seems to imply that $b^{ij}(\mathbf{x}) = 0$, it is not correct since, in general, it is not possible to exchange the order of the limit and the integral in Eq. (4). Therefore, it is possible that $b^{ij}(\mathbf{x})$ is not zero, even if $W(\mathbf{x}|\mathbf{y}, t) = 0$.

Deterministic process. If the diffusion matrix $b^{ij}(\mathbf{x})$ and the term $W(\mathbf{x}|\mathbf{y},t)$ vanish, then the Differential Chapman-Kolomogorov equation (2) is reduced to

$$\frac{\partial p(\mathbf{x},t)}{\partial t} = -\sum_{i=1}^{N} \frac{\partial}{\partial x^i}\{a^i(\mathbf{x})p(\mathbf{x},t)\}. \quad (6)$$

This equation is deterministic because it does not involve any random fluctuations. Moreover, this equation is essentially equivalent to the ordinary differential equation, and therefore, it is origin of many equations used frequently in biology. For example, the control theory used for analyzing chemical-taxi of E.coli in [20] is included in Eq. (6).

Diffusion process. In the absence of the jump term (i.e., $W(\mathbf{x}|\mathbf{y},t) = 0$), the Differential Chapman-Kolmogorov equation reads as

$$\frac{\partial p(\mathbf{x},t)}{\partial t} = -\sum_{i=1}^{N} \frac{\partial}{\partial x^i}\{a^i(\mathbf{x})p(\mathbf{x},t)\} + \frac{1}{2}\sum_{i,j=1}^{N} \frac{\partial^2}{\partial x^i x^j}\{b^{ij}(\mathbf{x})p(\mathbf{x},t)\}, \quad (7)$$

which is known as a diffusion process, which will be used later.

Jump process. In contrast, if we assume that $a^i(\mathbf{x}) = b^{ij}(\mathbf{x}) = 0$, the Differential Chapman-Kolmogorov equation takes the form

$$\frac{\partial p(\mathbf{x},t)}{\partial t} = \int dy[W(\mathbf{x}|\mathbf{y},t)p(\mathbf{y},t) - W(\mathbf{y}|\mathbf{x},t)p(\mathbf{x},t)], \quad (8)$$

which is known as a jump process. It means that the path trajectory of $\mathbf{X}_t$ will exhibit discontinuities (large jumps) at specific discrete points.

It is also known that the jump processes can represent some kind of chemical reactions (Ref. [16]). Therefore, we may use this equation to analyze the metabolic pathways [21] in cells, which are composed of chemical reactions and chemical compounds.

How to use the Differential Chapman-Kolmogorov equation. We can obtain the dynamics of probability density $p(\mathbf{x}, t \mid \mathbf{x}_0, t_0)$ for any time t, from experimental data of instantaneous transition probability $T_\epsilon(\mathbf{y}, \mathbf{x})$ (ϵ is sufficiently small and fixed), by the following procedure:

(i) Given the experimental data of instantaneous transition probability $T_\epsilon(\mathbf{y}, \mathbf{x})$ (ϵ is sufficiently small), we obtain $a^i(\mathbf{x})$, $b^{ij}(\mathbf{x})$ and $W(\mathbf{x}|\mathbf{y}, t)$ by using Eq. (3), (4) and (5) respectively.[2]

(ii) By inserting $a^i(\mathbf{x})$, $b^{ij}(\mathbf{x})$ and $W(\mathbf{x}|\mathbf{y}, t)$ into the Differential Chapman-Kolmogorov equation Eq. (2) and by solving this PDE (Partial Differential Equation), we can obtain useful information like the distribution, the expectation value, the variance and the correlation at any time.

2.1.2 Initial instantaneous transition data $T_\epsilon(\mathbf{y}, \mathbf{x})$

The initial instantaneous transition data $T_\epsilon(\mathbf{y}, \mathbf{x})$ of N genes for studying correlations should be determined by the experiment, which measures the short-time transition probability between the N dimensional gene expression levels $\mathbf{x}$ at time t and the N dimensional gene expression levels $\mathbf{y}$ at time $t+\epsilon$. Although some experiments have been done for measuring gene expression time series of many organisms [8, 15, 14], the statistics are not enough to completely determine the $T_\epsilon(\mathbf{y}, \mathbf{x})$ and we believe that several experiments should be done under the same condition to have enough statistics. Moreover, the stochastic nature of the fluctuations of the gene expression level, strongly supports the idea of *many experiments-many genes* under the same conditions.

Therefore, for the time being, we assume that the expression of the initial instantaneous transition data $T_\epsilon(\mathbf{y}, \mathbf{x})$ of N genes correlation is the Gaussian type, which seems general enough to illustrate our model. The expression is as follows;

$$\begin{aligned} & T_\epsilon(\mathbf{y}, \mathbf{x}) \\ = \; & \frac{1}{\sqrt{(2\pi\epsilon)^N det(\sigma^{ij})}} \\ & \exp\Big[-\frac{1}{2}\epsilon^{-1}\sigma_{ij}(y^i - x^i - \epsilon\mu^i(m^i - x^i))(y^j - x^j - \epsilon\mu^j(m^j - x^j)\Big], \end{aligned} \tag{9}$$

[2]Notice that we do not need the whole data $T_\epsilon(\mathbf{y}, \mathbf{x})$ for any ϵ. The necessary data is $T_\epsilon(\mathbf{y}, \mathbf{x})$ at a sufficiently small fixed ϵ.

where σ_{ij} is the inverse matrix of σ^{ij} (i.e. $\sum_k \sigma^{ik}\sigma_{kj} = \delta^i_j$). We note that the contraction of the indexes i, j is done by using the Einstein summation convention.

This ITP $T_\epsilon(\mathbf{y}, \mathbf{x})$ (Eq. (9)) allows us to analyze the gene correlation phenomena and, furthermore, has the mean reverting property, which is assumed in order to be compatible with the observation in [15, 14] (see Fig. 1). Here, we give some annotations on the parameters of $T_\epsilon(\mathbf{y}, \mathbf{x})$. m^i denotes the average expression level of genes i and μ^i means the tendency to reverting to m^i. Finally, σ_{ij} indicates the correlation between gene i and j. These parameters will be clear to the readers in the following sections.

2.1.3 Computation of $a^i(\mathbf{x})$, $b^{ij}(\mathbf{x})$ and $W(\mathbf{x}|\mathbf{y}, t)$

In this section, we compute $a^i(\mathbf{x})$, $b^{ij}(\mathbf{x})$ and $W(\mathbf{x}|\mathbf{y}, t)$ from the initial data of ITP $T_\epsilon(\mathbf{y}, \mathbf{x})$. Inserting Eq. (9) into (3), (4) and (5), we obtain

$$a^i(\mathbf{x}) = \lim_{\epsilon \to 0} \frac{1}{\epsilon} \int_{|\mathbf{y}-\mathbf{x}|<\delta} (y^i - x^i) T_\epsilon(\mathbf{y}, \mathbf{x}) d\mathbf{y} = \mu^i(m^i - x^i), \tag{10}$$

and

$$b^{ij}(\mathbf{x}) = \lim_{\epsilon \to 0} \frac{1}{\epsilon} \int_{|\mathbf{y}-\mathbf{x}|<\delta} (y^i - x^i)(y^j - x^j) T_\epsilon(\mathbf{y}, \mathbf{x}) d\mathbf{y} = \sigma^{ij}, \tag{11}$$

and

$$W(\mathbf{y}|\mathbf{x}, t) = \lim_{\epsilon \to 0} T_\epsilon(\mathbf{y}, \mathbf{x})/\epsilon = 0. \tag{12}$$

Here, we remark the following. The drift term $a^i(\mathbf{x}) = \mu^i(m^i - x^i)$ represents that our model has the mean value property (i.e. the gene expression level of gene i tends to revert the mean value m^i with the quickness μ^i. see Fig. 1). The diffusion matrix $b^{ij}(\mathbf{x}) = \sigma^{ij}$ denotes the correlation between gene i and j. Here, we remark that the correlation is constant in our model (Eq. (11)) for simplicity, although we could include the $\mathbf{x}$ dependence in the model in future work. Finally, the jump term $W(\mathbf{x}|\mathbf{y}, t) = 0$ means that our model does not contain the jump process.

2.1.4 Emergence of Kolmogorov equation and SPDE

In the last section, we find out that $a^i(\mathbf{x}) = \mu^i(m^i - x^i)$, $b^{ij}(\mathbf{x}) = \sigma^{ij}$ and $W(\mathbf{x}|\mathbf{y}, t) = 0$. However, in order to keep the argument more general, we still consider the drift term $a^i(\mathbf{x})$ and diffusion term $b^{ij}(\mathbf{x})$ arbitrary while we assume that the jump term vanishes $W(\mathbf{x}|\mathbf{y}, t) = 0$.

Kolmogorov Equation. If $W(\mathbf{x}|\mathbf{y},t) = 0$ vanishes, then the Differential Chapman-Kolmogorov equation (2) becomes the Kolmogorov equation:

$$\frac{\partial p(\mathbf{x},t)}{\partial t} = -\sum_{i=1}^{N} \frac{\partial}{\partial x^i}\{a^i(\mathbf{x})p(\mathbf{x},t)\} + \frac{1}{2}\sum_{i,j=1}^{N} \frac{\partial^2}{\partial x^i x^j}\{b^{ij}(\mathbf{x})p(\mathbf{x},t)\}, \quad (13)$$

where $p(\mathbf{x},t) = p(\mathbf{x},t|\mathbf{x}_0,t_0)$.

SPDE. In addition, it is known that the Kolmogorov equation is equivalent to the following stochastic partial differential equation (SPDE):

$$dX_t^i = \alpha^i(\mathbf{X}_t)dt + \sum_{j=1}^{N} \beta^{ij}(\mathbf{X}_t)dW_j(t). \quad (14)$$

Here, the multi-dimensional stochastic variable $\mathbf{X}_t$ denotes the gene expression level, $\alpha^i(\mathbf{x}) = a^i(\mathbf{x})$ denotes the average change of the instantaneous transition of the gene expression level per unit time, $\beta^{ij}(\mathbf{x})$ denotes the covariance of instantaneous transition of the gene expression level per unit time given by $b^{ij}(\mathbf{x}) = \sum_{k=1}^{N} \beta^{ik}(\mathbf{x})\beta^{jk}(\mathbf{x})$ (Here we remark that $\beta^{ij}(x)$ is not uniquely determined from $b_{ij}(x)$ due to the rotation group ambiguity.) and $\mathbf{W}(t) = (W_1(t), \cdots, W_N(t))$ denotes the multi-dimensional Wiener process where all the processes are independent of each other $dW_i(t)dW_j(t) = \delta_{ij}dt$. We also note that the stochastic calculus in our approach follows the Ito rule by construction, and not the Stratonovich rule [3].

Correlated-SPDE. Although the above SPDE (14) is good for a theoretical study, it is difficult to analyze Eq. (14) directly, since Eq. (14) includes the multi-components processes of $W_1(t), \cdots, W_N(t)$ and the ambiguity of $\beta^{ij}(\mathbf{x})$ relative to the rotation group. Therefore, we transform the above SPDE (14) into a more convenient expression (i.e, where the correlation is explicitly manifested) as follows:

$$dX_t^i = \alpha^i(\mathbf{X}_t)dt + \beta^i(\mathbf{X}_t)dz^i(t), \quad (15)$$

[3] See [16, 18] for further discussions on the Ito-Stratonovich dilemma.

and the correlation is given by

$$dz^i(t)dz^j(t) = \rho^{ij}(x)dt, \tag{16}$$

where $\beta^i(x) = \sqrt{b^{ii}(x)}$, $\rho^{ij}(x) = \frac{b^{ij}(x)}{\sqrt{b^{ii}(x)}\sqrt{b^{jj}(x)}}$ and $(-1 \leq \rho_{ij}(x) \leq 1)$.

It is important to note that this expression is more easy to understand and deal with from the practical point of view, since the SPDE (15) has only one Brownian motion (fluctuations) component $dz^i(t)$, while many components of Brownian motion appear as $dW^1(t).\cdots, dW^N(t)$ in (14). Furthermore, all the correlation information among genes are gathered in $dz^i(t)dz^j(t) = \rho^{ij}(\mathbf{x})dt$ in Eq. (16), which by the way is a more easy expression to deal with.

2.1.5 Analysis of our model

¿From general analysis of Kolmogorov equation, we return to our original situation where the drift $a^i(\mathbf{x}) = \mu^i(m^i - x^i)$, the diffusion $b^{ij}(\mathbf{x}) = \sigma^{ij}$ and the jumping term $W(\mathbf{x}|\mathbf{y}, t) = 0$. Then, by inserting them into the SPDE Eq. (15) and (16), we obtain

$$dX_t^i = \mu^i(m^i - X_t^i)dt + \sigma^i dz_t^{(i)}, \tag{17}$$

where $\sigma^i = \sqrt{\sigma^{ii}}$ and the correlation is given by

$$dz^i(t)dz^j(t) = \rho^{ij}dt = \frac{\sigma^{ij}}{\sigma^i\sigma^j}dt. \tag{18}$$

This reduced model is also known as Vasicek model in financial engineering (see Ref. [19]). This SPDE directly gives useful information about the properties of the model as follows:

(i) Mean reverting. The model (Eq. (17)) has mean reverting property. We can observe this phenomena in Fig. 1, which is obtained from data of gene expression time series experiments of human and yeast organism [15, 14].

(ii) Multi-correlation. Eq. (17) also exhibits an embedded correlation relationship given by Eq. (18). Therefore, by using this relationship, we can analyze the gene correlation phenomena among genes using our model.

2.1.6 Gene expression dynamical solution

By using Ito formula, we can solve the SPDE (Eq. (17)) and derive the dynamical solution of gene expression

$$X_t^i = m^i + (x_0^i - m^i)e^{-\mu^i t} + \sigma^i \int_0^t e^{-\mu^i(t-s)} dz^i(s), \tag{19}$$

where $x_0^{(i)} = X_0^{(i)}$ is the initial value of gene expression level. ¿From this solution, we can obtain the most important quantities which contain the relevant information of the system: expectation value and variance and covariance of multi-dimensional gene expression level at any time as follows:

$$E[X_t^i] = m^i + (x_0^i - m^i)e^{-\mu^i t}, \tag{20}$$

$$V[X_t^i] = \frac{(\sigma^i)^2}{2\mu^i}(1 - e^{-2\mu^i t}), \tag{21}$$

and

$$Cor[X_t^i, X_t^j] = \frac{\sigma^i \sigma^j \rho^{ij}}{\mu^i + \mu^j}(1 - e^{-(\mu^i + \mu^j)t}). \tag{22}$$

Roughly speaking, it means that if we specify the initial state ($x_0^{(i)} = X_0^{(i)}$) (for example, a patient receives a chemical treatment which modifies the amount of mRNA in cells), then we may predict the effect of that treatment in cells by knowing the expectation value, variance and covariance of multi-dimensional gene expression level at any time.

2.2 Computation of simulated data

In order to obtain the sample path of multi-gene correlation dynamics from Eq. (17), we use the difference equation corresponding to Eq. (17) as follows. For sufficiently large number n. We equally divide the time interval $[0, t]$ by $t_i = i\Delta t$ ($i = 0, \cdots, n$ and $\Delta t = t/n$), then from Eq. (17) we obtain:

$$X_{i+1}^j = X_i^j + \mu^j(m^j - X_i^j)dt + \sigma^j \Delta z_t^{(j)}, \tag{23}$$

where

$$X_i^j = X_{t_i}^j \qquad \text{and} \qquad \Delta z_t^{(j)} = z_{t_{i+1}}^{(j)} - z_{t_i}^{(j)} \sim N_n(0, \rho^{ij}), \tag{24}$$

where $N_n(0, \rho^{ij})$ denotes the n-dimensional normal distribution and the mean vector is zero and the variance matrix is given by ρ^{ij}. We repeatedly use Eq. (23) to obtain the sample path. In Fig. 2, we show the five sample paths for total positive correlation $\rho^{12} = 1$, slightly positive correlation $\rho^{12} = 0.5$, no correlation $\rho^{12} = 0$, slightly negative correlation $\rho^{12} = -0.5$ and total negative correlation $\rho^{12} = -1$, respectively.

Although the correlation ρ^{ij} between genes is easy to be characterized by the theoretical analysis of simulated data of gene expression (Eq. (9)), this issue becomes more difficult in a practical problem. Precisely, the most important experimental problem related to the analysis of time series of gene expression data by using Microarrays/GeneChips technologies is to identify significant correlations between observables of genes. The criteria that we may use for considering that two genes are significantly correlated is as follows: (1) By using a given time-series set of gene expression experimental data, our theory can be used to calculate the correlation value ρ^{ij} for each couple of genes. In the case that this value is in the vicinity of the value one ($\rho^{ij} \simeq 1$) for two genes, we may consider that both genes are significant enough correlated.

In addition, the following criterias should also be taken into account. (2) The correlation ρ^{ij} between genes in our approach is obtained by inserting the ITP $T_\epsilon(\mathbf{y}, \mathbf{x})$ into the model. This ITP $T_\epsilon(\mathbf{y}, \mathbf{x})$ is obtained by using experimental techniques as Microarry/GeneChips. Currently, these technologies have a non-zero instrumental noise that in some cases may exceed 30%. Therefore, it is important that this source of noise is reduced as much as possible for each experiment in order to evaluate with more accuracy the correlation between genes. Finally, (3) our theory for predicting the correlation phenomena is based on a stochastic approach. As we explained through the text, and in more extension in the next section, many experiments for analysing gene expression time series are encouraged to have a enough statistics to achieve a precise interpretation of the dynamics of gene correlation. Therefore, by increasing the number of the experiments, the confidence of the correlation observable ρ^{ij} predicted by our theory would be improved.

3 Experimental proposal

The most important factor in our model is the initial data of ITP $T_\epsilon(\mathbf{y}, \mathbf{x})$, which characterizes the gene correlation system. However, as far as we know,

we do not have enough experimental data for completely determining the ITP $T_\epsilon(\mathbf{y}, \mathbf{x})$. Therefore, to determine this ITP $T_\epsilon(\mathbf{y}, \mathbf{x})$, we propose an experiment, which measures the short-time transition probability between the N dimensional gene expression levels $\mathbf{x}$ at time t and the N dimensional gene expression levels $\mathbf{y}$ at time $t + \epsilon$. The novelty, is that such experiments should be carried out at least hundred times to have enough statistics, and in addition, should be done under the same external condition.

Interestingly, the stochastic nature of the fluctuations of the gene expression level [9], strongly supports the idea of *many experiments-many genes* under the same conditions.

If we can obtain the experimental data of $T_\epsilon(\mathbf{y}, \mathbf{x})$, our model can predict the future behavior of multi-gene correlation using our construction (for example, the expectation value and variance of the gene expression level at any time in the future) from the initial value of gene expression levels. In particular, it would contribute to uncover how to regulate specific genes by applying some external action (e.g., a patient under medical treatment).

4 Conclusions

We have carried out a theoretical study on gene expression correlation, which is one of the crucial topics of genomics in the current post-sequence era. Our study indicates that it is possible to analyze the dynamics underlying the gene expression correlation phenomena by using only one assumption the Markov property. In other words, it means that *the multi-dimensional correlation dynamics of gene expression obeys the Markov property.*

Our theoretical approach of multi-gene expression dynamics indicates that we can specify an initial state ($\mathbf{X}_{t_0} = \mathbf{x}_0$) of gene expression in a cell and be able to predict the most relevant observables of the distribution of genes as expectation value, variance and covariance of multi-dimensional gene expression level at any time in the future. This feature represents an important step forward in the current analysis of gene correlation analysis and have potential implications for genetic engineering, for example by developing personalized medicines according to the features of individuals.

Furthermore, in order to achieve the above described goals we presented an experimental proposal. The main idea of this new proposal is that *many* experiments of *many* genes would be useful for completely uncover the dynamics of multi-genes in cells.

It is also worth noticing that stochastic theory offers a huge and rich variety of tools for studying the gene expression fluctuations, and by extension many other cellular phenomena. For example, it is known that the jump processes described by Eq. (8) can represent some kind of chemical reactions. Therefore, as future work we may use that equation to analyze the metabolic pathways in cells, which are composed of chemical reactions and chemical compounds.

The availability of complete genomes for several organisms has definitely opened new and exciting possibilities of studying the gene correlation dynamics and mechanisms. Consequently, we believe that our theoretical model, together with the experimental proposal, may further serve to understand the regulatory interactions among genes and contribute to enlighten the advances of the post-sequence era.

References

[1] L.H. Hartwell, J.J. Hopfield, S. Leibler and A.W. Murray, Nature **402**, C47 (1999).

[2] D. J. Lockhard, E. A. Winzeler, Nature **405**, 827 (2000).

[3] T. Ochiai, J.C. Nacher, T. Akutsu, Physics Letters A **330**, 313 (2004).

[4] W. Li, Y. Yang, J. Theor. Biology **219**, 539 (2002).

[5] D.C. Hoyle, M. Rattray, R. Jupp, A. Brass, Bioinformatics **18**, 576 (2002).

[6] V.A. Kuznetsov, G.D. Knott, R.F. Bonner, Genetics **161**, 1321 (2002).

[7] C. Furusawa and K. Kaneko, Physical Review Letters **90**, 088102 (2003).

[8] H. Ueda *et. al.*, Proc. Natl. Acad. Sci. U.S.A. **101**, 3765 (2004).

[9] M.B. Elowitz, A.J. Levine, E.D. Siggia, P.E. Swain, Science **297**, 1183 (2002).

[10] J. Paulsson, Nature **427**, 415 (2004).

[11] W.J. Blake, M. Kaern, C.R. Cantor, J.J. Collins, Nature **422**, 633 (2003).

[12] J. Hasty, J. Pradines, M. Dolnik, J.J. Collins, Proc. Natl. Acad. Sci. U.S.A **97**, 1075 (2000).

[13] K. Sato, Y. Ito, T. Yomo, K. Kaneko, Proc. Natl. Acad. Sci. U.S.A **100**, 14086 (2003).

[14] R.J. Cho et al., Mol. Cell. Biol. **2**, 65 (1998).

[15] R.J. Cho et al., Nature Genetics **27**, 48 (2001).

[16] N.G.van Kampen, *Stochastic processes in physics and chemistry,* Elsevier Science B.V. (1992).

[17] E. Wong, *Stochastic Processes in Information and Dynamical Systems*, Ed. New York, McGraw-Hill (1971).

[18] C.W. Gardiner, *Handbook of Stochastic Methods for physics, chemistry and the natural sciences*, Springer Verlag (1990)

[19] T. Mikosch, *Elementary Stochastic Calculus with Finance in View*, World Scientific Publishing Co. Pte. Ltd (1998).

[20] T.-M. Yi, Y. Huang, M. I.Simon, J. Doyle, Proc. Natl. Acad. Sci. U.S.A **97**, 4649 (2000).

[21] M. Kanehisa, S. Goto, S. Kawashima, Y. Okuno, M. Hattori, Nucleic Acids Research, **32**, D277 (2004).

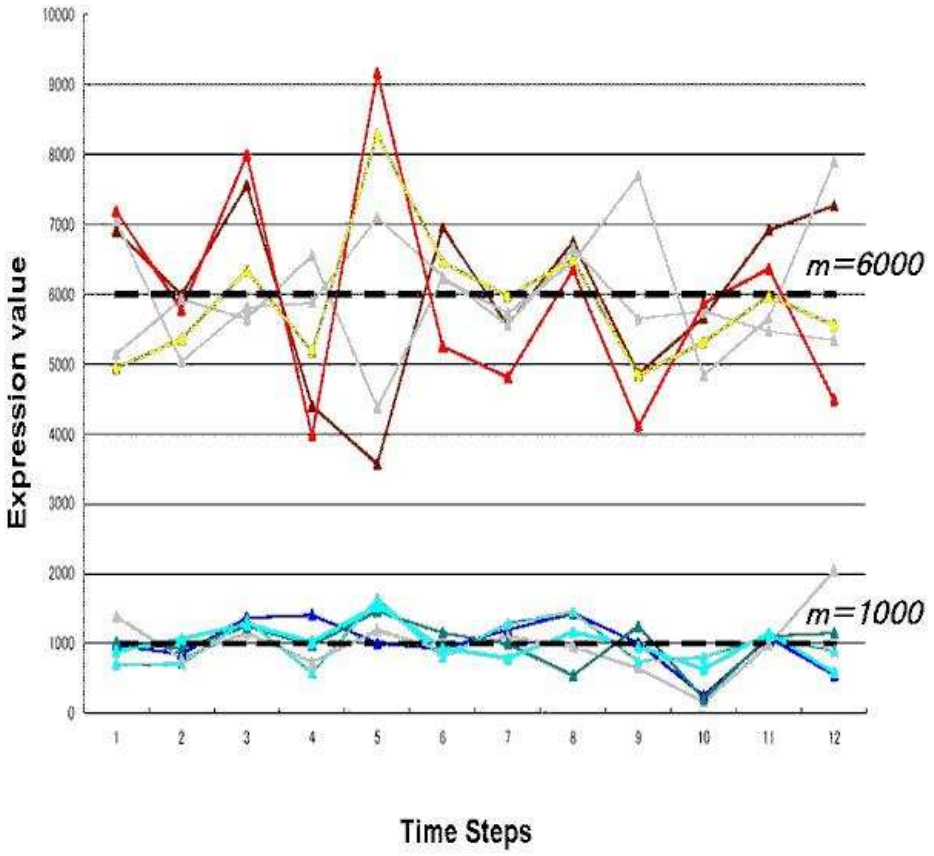

Figure 1: We show the experimental absolute value of gene expression level (vertical axis) vs. time (horizontal axis) of a selected group of genes which belong to human organism [15]. We see that the gene expression value fluctuates around the mean value m=6000 and m=1000.

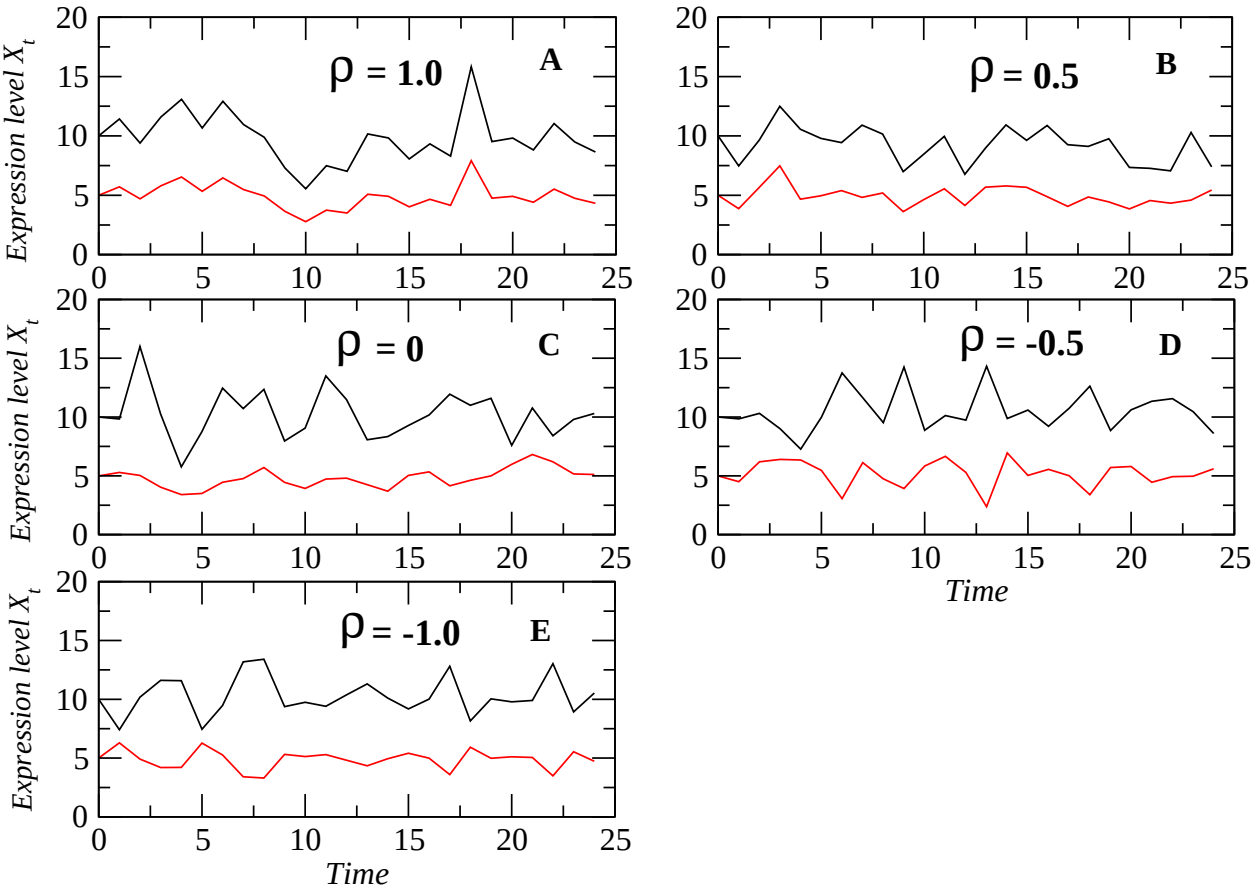

Figure 2: We show the simulated results of our model for different values of ρ. Values of ρ are indicated in each figure, from A to E. (A) $\rho = 1.0$ indicates that both genes are totally positively correlated. (B) $\rho = 0.5$ indicates that both genes are slightly positively correlated. (C) ρ=0 indicates uncorrelation between genes. (D) $\rho = -0.5$ indicates that both genes are slightly negatively correlated. (E) $\rho = -1.0$ indicates that both genes are completely negatively correlated.